
REACTION

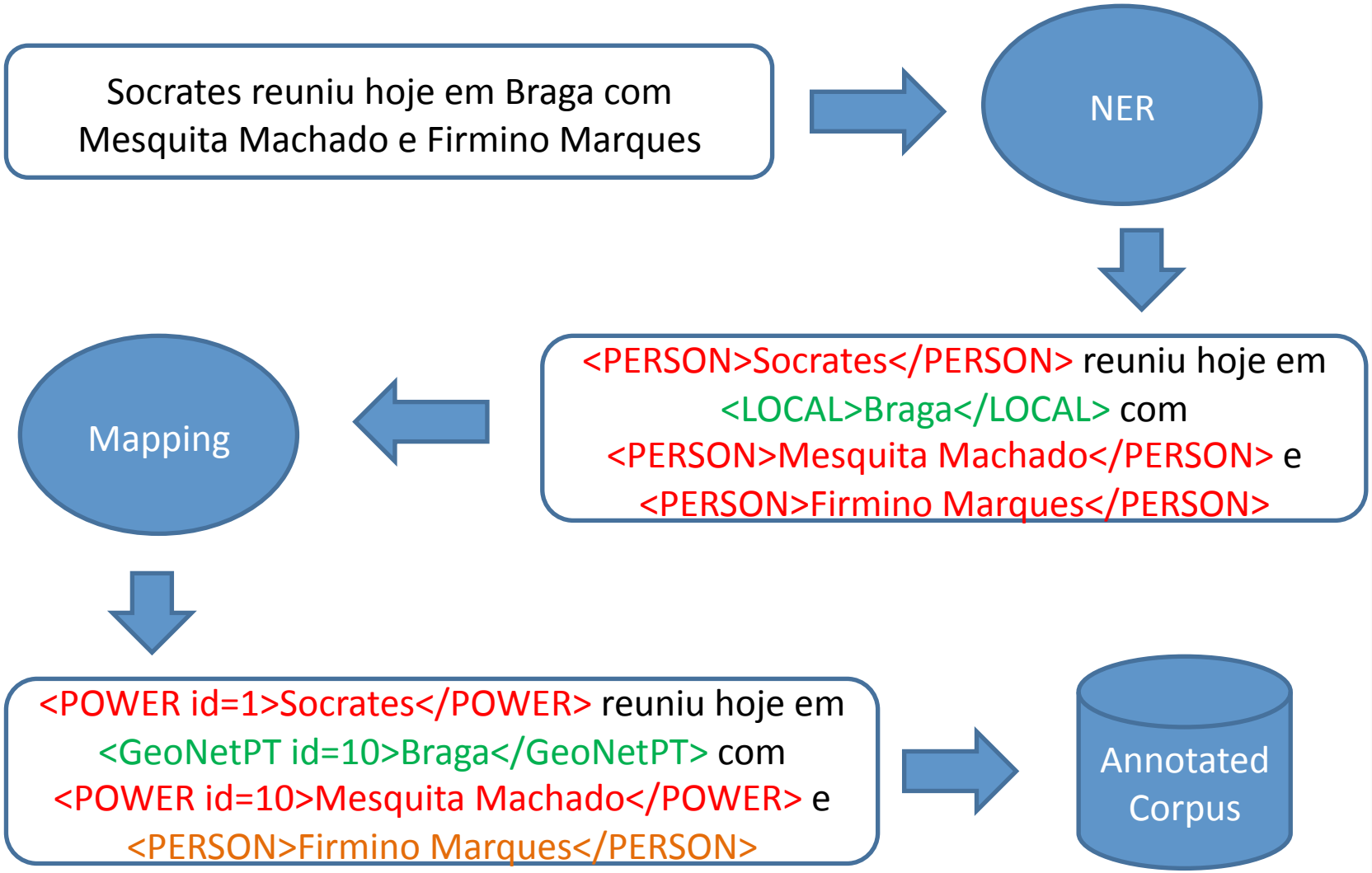
Information Discovery

Relationship extraction techniques to support **information discovery in journalists' activities**

- **Entity Ranking**: finding the relevant entities for a given topic
- **Entity Distillation**: finding relevant resources for a given entity
- **Attribute Selection**: finding a list of key aspects to compare and differentiate a given set of entities

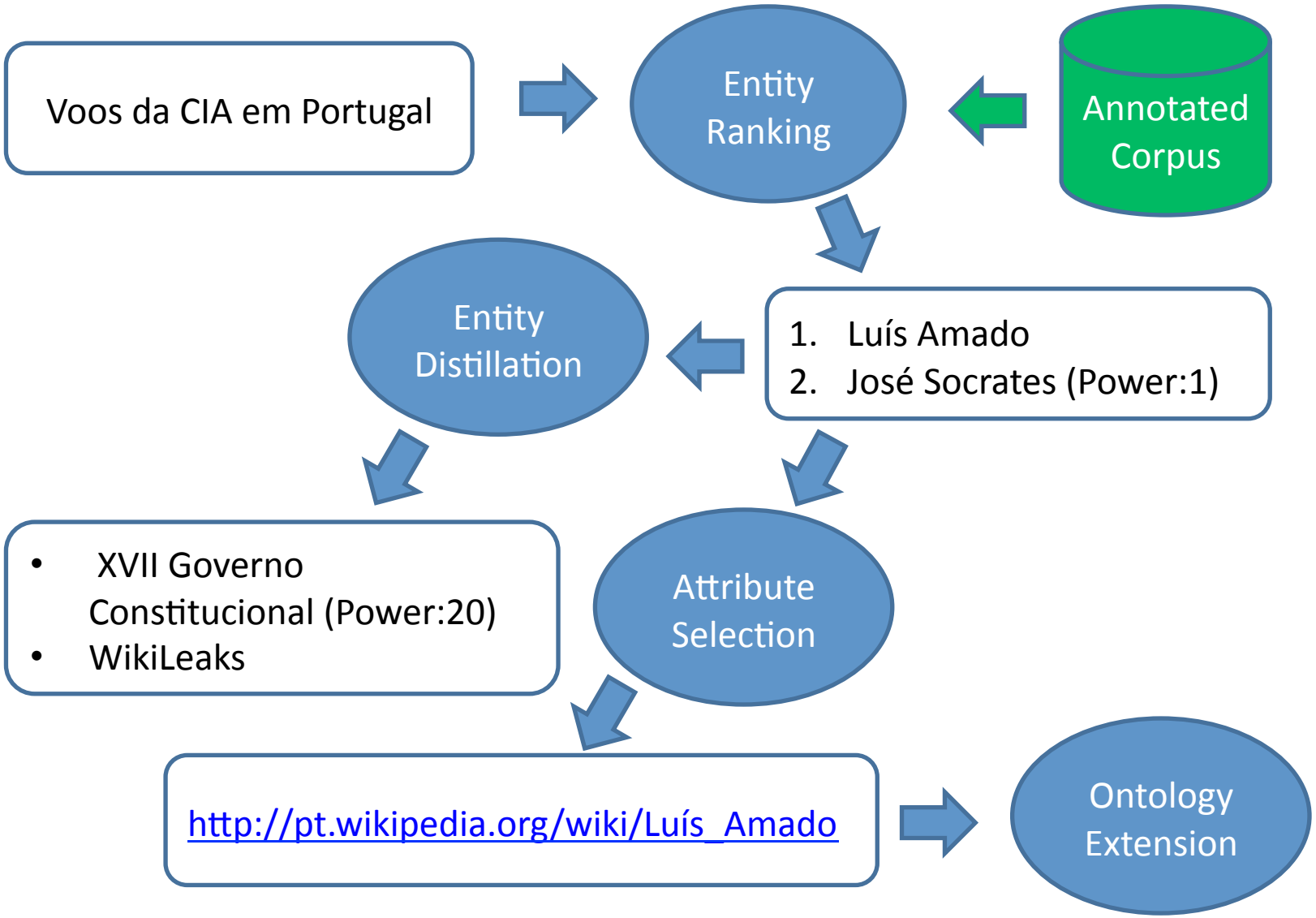
REACTION

Annotation



REACTION

Analysis



REACTION

Input Corpus

- crawling portuguese news from publico.pt:
  - **economy**
  - **education**
  - **politics**
  - **society**
  - **local**
- Working with last 6 months
- But 10 years available

REACTION

Prototype

- NER
  - Rembrandt
    - Several types of entities
- Mapping (Grounding)
  - string matching based on Evidence Content
  - Weighted Jaccard Similarity

$$Sim_{PM} = \frac{\sum_{w \in (d1 \cap d2)} EC(w)}{\sum_{w \in (d1 \cup d2)} EC(w)}$$

- Output XML

REACTION

Ontologies in Mapping

- **POWER** Politics Ontology for Web Entity Retrieval
- **Yahoo!GeoPlanet** world-wide geographic ontology
- **Geo-Net-PT** geographic ontology covering Portugal

REACTION

Next Steps

- Build an API
- Improve NER/Grounding
  - Machine Learning
  - Same features used for Power enrichment
- Participation in TREC and TAC

REACTION

TREC 2011 - Entity

- **Related entity finding (REF)**
- **Task**: return a ranked list of entities of a specified type that engage in a given relationship with a given source entity.
- **Collection**: ClueWeb09 English
- **Entity identification**: Homepage
- **Topics**: 50 new topics for 2011

REACTION

TREC Corpus

- **Web Pages:**
  - 1,040,809,705 web pages, in 10 languages
  - 5 TB, compressed. (25 TB, uncompressed.)
- **Web Graph:**
  - **Entire Dataset:**
    - Unique URLs: 4,780,950,903 (325 GB uncompressed, 105 GB compressed)
    - Total Outlinks: 7,944,351,835 (71 GB uncompressed, 24 GB compressed)
- **TREC Category B (first 50 million English pages)**
  - Unique URLs: 428,136,613 (30 GB uncompressed, 10 GB compressed)
  - Total Outlinks: 454,075,638 (3 GB uncompressed, 1 GB compressed)

REACTION

Text Analysis Conference (TAC) 2011

- Knowledge Base Population (KBP2011) Track
- "...discover information about named entities and incorporate this information in a knowledge source"
- knowledge base:
  - derived from Wikipedia infoboxes
- collection of documents:
  - mostly news articles
- Entity Linking:
  - given an entity name, return the identifier of the entity in the KB or NIL if doesn't exist.
- Slot Filling:
  - given an entity name and it's type, a list of attributes and optional an ID of the entity in the KB, discover the attributes of the specified entity from the document collection and expand the KB.

REACTION